People Detection in Color and Infrared Video using HOG and Linear SVM

Pablo Tribaldos[1], Juan Serrano-Cuerda[1], María T. López[1,2], Antonio Fernández-Caballero[1,2], and Roberto J. López-Sastre[3]

[1] Instituto de Investigación en Informática de Albacete (I3A), 02071-Albacete, Spain
[2] Universidad de Castilla-La Mancha, Departamento de Sistemas Informáticos, 02071-Albacete, Spain
Antonio.Fdez@uclm.es
[3] Universidad de Alcalá, Dpto. de Teoría de la señal y Comunicaciones, 28805-Alcalá de Henares (Madrid), Spain

Abstract. This paper introduces a solution for detecting humans in smart spaces through computer vision. The approach is valid both for images in visible and infrared spectra. Histogram of oriented gradients (HOG) is used for feature extraction in the human detection process, whilst linear support vector machines (SVM) are used for human classification. A set of tests is conducted to find the classifiers which optimize recall in the detection of persons in visible video sequences. Then, the same classifiers are used to detect people in infrared video sequences obtaining excellent results.

Keywords: Human classification, Color video, Infrared video, HOG, Linear SVM

1 Introduction

In smart spaces visual surveillance, real-time detection of people (e.g. [1], [2]) and their activities [3] is performed both in visible (e.g. [4], [5], [6]) and infrared spectrum (e.g. [7], [8]). Therefore, it seems interesting to find a single solution to detect people in both types of videos. Most methods described for the detection of people are divided into two steps, namely extraction of image features and classification of the images according to these features.

In this sense, histogram of oriented gradients (HOG) is a feature extraction technique for the detection of objects [9]. Its essence is that the shape of an object in an image can be described by means of the intensity distribution of the gradients. The great advantage of a detector obtained using HOG descriptors is that it is invariant to rotation, translation, scaling and illumination changes. Therefore, it has been applied successfully in both visible spectrum images (e.g. [10], [11], [12], [13]) and infrared images (e.g. [14], [15], [16], [17]). In our approach, we are firstly interested in discovering if there are HOG descriptors for extracting human features that are equally valid for color and infrared images.

After using HOG descriptors, support vector machines (SVM) are usually used in the classification stage. SVM are a set of supervised learning algorithms which were introduced for linearly separable [18] and linearly non-separable [19] data. SVM have been used in classification and regression problems in many fields such as text recognition, bioinformatics and object recognition, among others. They have also been used successfully in the detection of persons (e.g. [20], [21]). Here, we are also interested in knowing if linear SVM trained with color images provide good results in classifying infrared images without re-training. Should this be true, we could overcome the lack of large enough datasets in the infrared spectrum.

2 Detection of Humans in Color and Infrared Video

2.1 HOG for Feature Extraction

Histogram of oriented gradients (HOG) consists of a series of steps that provide an array of image features representing the objects contained in an image in a schematic manner. The image features are later used to detect the same objects in other images. In our particular case, we are interested in obtaining strong features for human detection.

Global normalization of the gamma/color image. This first step is undertaken to reduce the influence of the effects of image lightning changes. In order to normalize the color of an image, histogram equalization is applied. The $\sqrt{RGB}$ function is used for gamma normalization. Each pixel is obtained from the square root of its channel values.

Gradient computation. A first derivative edge detection operator is launched to estimate the image gradients. Specifically, filter kernels $G_x = [-1\ 0\ 1]$ and $G_y = [-1\ 0\ 1]^T$ are applied to x and y axes, respectively, as well as a smoothing value $\sigma = 0$. This way the image contours, shape and texture information are obtained. Furthermore, resistance to illumination changes is achieved. The gradient is calculated for each color channel, and the locally dominant gradient is used to achieve invariance against color.

Orientation binning. This step generates the HOG descriptors. Local information on the direction of the gradient is used in the way SIFT [22] does. It aims to produce an encoding that is sensitive to the local image content, while being resistant to small changes in attitude or appearance. Orientation binning divides the image into regions called "cells" of $n \times n$ pixels. Gradients or orientations of the edges at each cell pixels are accumulated in a 1-D histogram. The combined histograms form the orientation histogram. Each orientation histogram divides the range of angles of the gradient in a fixed number of bins. The gradient value of each pixel of the cell is used in the orientation histogram for voting.

Local normalization. Now, the cells are grouped into sets called "blocks", and each cell block is normalized. A cell can belong to several overlapping blocks. Therefore, it appears several times in the final vector, but with different normalization. Indeed, the normalization of each block depends on the cell which it belongs to. Normalization provides better invariance against lightning, shadows and contrast of the edges. The descriptors of the normalized blocks are precisely the HOG descriptors. Dalal and Triggs [9] explore four different methods for block normalization: L1-norm (see equation (1)), L1-sqrt (2), L2-norm (3) and L2-hys. Let ν be the non-normalized vector containing all histograms in a given block, $\|\nu\|_k$, its k-norm for $k = 1, 2$ and e be some small constant (the exact value, hopefully, is unimportant). Finally L2-hys is L2-norm followed by clipping (limiting the maximum values of ν to 0.2) and re-normalizing. The normalization factor can be one of the following:

- L1-norm

$$f = \frac{\nu}{(\|\nu\|_1 + e)} \tag{1}$$

- L1-sqrt

$$f = \sqrt{\frac{\nu}{(\|\nu\|_1 + e)}} \tag{2}$$

- L2-norm

$$f = \frac{\nu}{\sqrt{\|\nu\|_2^2 + e^2}} \tag{3}$$

HOG descriptors combination. In the last stage of the process all blocks are combined into a dense grid of overlapping blocks, covering the detection window to obtain the final feature vector.

2.2 Linear SVM for Classification

Given the features of two objects, an SVM seeks a hyperplane optimally separating the features of an object from the other. An SVM maximizes the margin of separation between the two classes, so that one side of the hyperplane contains all objects of a class, and the other one the other objects. The vectors closest to the margin of separation are called support vectors and are used for classification. The accuracy of an SVM may be degraded in the case that data are not normalized. Normalization can be performed at the level of input features or at kernel level (in the feature space).

The classification task involves separating data into training and testing. Each instance of the training set contains a target value, which is the class label, and a series of attributes such as the observed features. The goal of SVMs is to create a model based on training data to predict the target values of the test dataset by only knowing their attributes. Given a training set with instance-label pairs $(x_i, y_i), i = 1, \ldots, l$, where $x_i \in R^n$ e $y \in \{0, -1\}^l$, an SVM requires the solution of the following optimization problem:

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{4}$$

subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{5}$$

Here training vectors x_i are mapped into a large or even infinite dimensional space by function ϕ. SVMs seek a linear hyperplane with the maximum margin separator in this dimensional space. $C > 0$ is the error penalty parameter. Function $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function. LibSVM [23] offers the following four main kernel types:

- linear: $K(x_i, x_j) = x_i^T x_j$.
- polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$.
- radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \gamma > 0$. Variable γ can be expressed as $\gamma = 1/(2\sigma^2)$.
- sigmoidal: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$.

In this work it was decided to use a linear kernel, $K(x_i, x_j) = x_i \times x_j$, where $x_i, x_j \in N$ are the feature vectors. A linear kernel uses to work fine when handling only two classes and is quite easier to refine, as it only has one parameter affecting performance, namely C, the soft margin constant.

3 Data and Results

3.1 Parameters for Performance Evaluation

Let us define *positive image* as an image containing one or more persons and *negative image* as an image where no person appears. The parameters used to validate the goodness of the proposed classifier are:

- FP (false positives): number of images that are negative but have been classified as positive.
- FN (false negatives): number of images that are positive but have been classified as negative.
- TP (true positives): number of positive images that are correctly classified, that is, number of hits.
- P: number of positive test images.
- N: number of negative test images.
- T: number of test images:

$$T = P + N \tag{6}$$

- *accuracy*: percentage of the number of correctly classified test images:

$$accuracy = \frac{TP}{P} \cdot 100 \tag{7}$$

– *precision*: percentage of true positives among all positives detected:

$$precision = \frac{TP}{TP + FP} \cdot 100 \tag{8}$$

– *recall*: percentage of true positives among all positives:

$$recall = \frac{TP}{TP + FN} \cdot 100 \tag{9}$$

3.2 Parameters for HOG Feature Extraction

The recommended parameters used for extracting HOG descriptors [9] are provided in Table 1. These have been used without modifications in our approach.

Table 1. Recommended values for the extraction of HOG features.

Parameter	Value
Window size	64×128 pixels
Block size	2×2 cells
Cell size	8×8 pixels
Number of angle divisions	9 (no sign, 180°)
Overlap	8×8 pixels (stride = 8)
Gaussian smoothing	No
Histogram normalization	L2-hys
Gamma correction	Yes
Max number of detection window scalings	64

3.3 People Detection in Color Video

Description of training and test databases. Two people image databases widely addressed in the scientific community have been used to train and test the proposal in the visible spectrum. These are INRIA (Institut National de Recherche en Informatique et en Automatique) "Person Dataset" (available at `http://pascal.inrialpes.fr/data/human/`) and MIT (Massachusetts Institute of Technology) "Pedestrian Data" (available at `http://cbcl.mit.edu/software-datasets/PedestrianData.html`). The MIT training database of people was generated from color images and video sequences taken in a variety of seasons using several different digital cameras and video recorders. The pose of the people in this dataset is limited to frontal and rear views. The MIT pedestrian database contains 923 positive images; each image was extracted from raw data and was scaled to the size 64×128 and aligned so that the person's body was in the center of the image. The data is presented without any normalization.

The INRIA person database also contains images of people in different positions, backgrounds and with different lightning (see Table 2). There are also people partially occluded. This dataset was collected as part of research work on detection of upright people in images and video. The dataset contains images from several different sources. Only upright persons (with person height > 100)

Table 2. Description of INRIA person database.

	# of images	Size (pixels)
Positive training images	2,416	96×160
Negative training images	1,218	Not normalized
Positive test images	1,126	70×134
Negative test images	453	Not normalized

Table 3. Description of final test dataset.

	# of images	Size (pixels)
Positive training images	4262	64×128
Negative training images	12180	64×128
Positive test images	1126	64×128
Negative test images	4530	64×128

are marked in each image, and annotations may not be right; in particular at times portions of annotated bounding boxes may be outside or inside the object.

Before using images from this database to extract their features, it is recommended to normalize them to 64×128 pixels, and to get sub-images of their negative images. In our case, we have added:

- 10 sub-images of size 64×128 pixels are randomly extracted from each negative image.
- A centered window of size 64×128 pixels is extracted from each positive image.
- A mirror image of each positive image (reflection on the vertical axis) is obtained.

Therefore, $2,416$ positive training images, $12,180$ negative training images, $1,126$ positive test images and $4,530$ negative test images are extracted. Also, in order to increase the number of positive images to train the classifier, the mirror images of the MIT dataset are obtained. Table 3 shows the final set of images used for training and testing.

Description of the training process. During the training process the SVMs supplied by LibSVM and LibLINEAR [25] are used. The models generated will be used later for human detection. LibLINEAR offers several SVMs for linear classification; we use L2-regularized L1-loss (dual), L2-regularized L2-loss (primal) and L2-regularized L2-loss (dual). The influence of the soft margin constant C on the three kernels is studied. Table 4, Table 5 and Table 6 show the results for each kernel, respectively.

The aim is to find the best kernels to classify the color training input images to apply them to the detection of people in new images. The more accurate the results of the kernel are, the better the future detection results. In this case, in order to assess the goodness of a kernel we will use the *recall* evaluation parameter to obtain the minimum possible number of false negatives, although some more false positives may appear. From the previous study, we conclude to use kernels L2-regularized L2-loss (dual) with $C = 10$ and L2-regularized L2-loss (dual) with $C = 0.001$, as their respective *recall* values are very close (96.36% and $96,63\%$).

Table 4. Influence of parameter C in the L2-regularized L2-loss (dual) linear kernel.

C	Hits	FP	FN	precision (%)	recall (%)	accuracy (%)
0.0001	5582	16	58	98.57904085	94.849023	98.69165488
0.001	5590	24	42	97.86856128	96.269982	98.8330976
0.1	5580	37	39	96.71403197	96.536412	98,6562942
1	5574	41	41	96.35879218	96.358792	98.55021216
10	5577	41	38	96.35879218	96.625222	98.60325318
100	5577	41	38	96.35879218	96.625222	98.60325318

Table 5. Influence of parameter C in the L2-regularized L2-loss (primal) linear kernel.

C	Hits	FP	FN	precision (%)	recall (%)	accuracy (%)
0.0001	5581	16	59	98.57904085	94.760213	98.67397454
0.001	5591	24	41	97.86856128	96.358792	98.85077793
0.1	5577	36	43	96.80284192	96.181172	98.60325318
1	5578	35	43	96.89165187	96.181172	98.62093352
10	5578	36	42	96.80284192	96.269982	98.62093352
100	5578	36	42	96.80284192	96.269982	98.62093352

Table 6. Influence of parameter C in the L2-regularized L1-loss (dual) linear kernel.

C	Hits	FP	FN	precision (%)	recall (%)	accuracy (%)
0.0001	5565	21	70	98.13499112	93.783304	98.39108911
0.001	5592	16	48	98.57904085	95.737123	98.86845827
0.1	5578	38	40	96.62522202	96.447602	98.62093352
1	5575	43	38	96.18117229	96.625222	98.5678925
10	5576	42	38	96.26998224	96.625222	98.58557284
100	5577	41	38	96.35879218	96.625222	98.60325318

Description of the results. The classification in the visible spectrum is performed on the test images by using both selected kernels. Here, the best *accuracy* is 90.33% using the classification model L2-regularized L2-loss (dual) [24] with $C = 10$. The performance results are offered in Table 7. Also, some result images in the visible spectrum are shown in Fig. 1.

3.4 People Classification in Infrared Video

In order to test the proposal in infrared spectrum, we manually labeled 112 infrared images recorded by our research team. Of course, as stated previously, we use the parameters and kernels obtained during the training color images detection and classification phases. Here, the best performance results obtained for human detection in infrared are offered in Table 8. These come from the use of model L2-regularized L2-loss (primal) with $C = 0.001$. Now, *accuracy* is 94.64, which is astonishingly very high compared to the *accuracy* in color images. The reason for this increment is probably the fact the infrared images have been annotated manually and very carefully. Lastly, some resulting images in infrared are shown in Fig. 2.

4 Conclusions

The initial objective of this work was to efficiently detect humans in color and infrared video. For this, we use the HOG algorithm for extracting image features,

Table 7. Performance results in color video.

Kernel	L2-regularized L2-loss (primal) C= 0.001	L2-regularized L2-loss (dual) C = 10
Mean detection time (ms)	336	297
Hits	896	908
False positives	63	32
False negatives	222	229
Accuracy (%)	89.41	90.33
Precision (%)	93.43	96.60
Recall (%)	80.14	79.86

Fig. 1. Some results in color video.

and a linear SVM for classification of the features. This combination allows detecting humans in images with high accuracy, both in visible and infrared spectrum. The HOG algorithm obtains the feature vectors from the training color images of the proposed databases. Then, a linear SVM seeks a hyperplane capable of separating the feature vectors in two classes in the most optimal way.

As it has been demonstrated, after using the recommended parameters for the feature detector and selecting a couple of kernels suited for SVM in the color spectrum, the approach works well for in the visible and infrared spectra, providing an *accuracy* of 90.33% and a *recall* of 79.86% for automatically annotated images in the visible spectrum, and an *accuracy* of 94.64% and *recall* of 96.91% for manually annotated infrared images.

Acknowledgements

This work was partially supported by Spanish Ministerio de Economía y Competitividad / FEDER under TIN2010-20845-C03-01 and TIN2010-20845-C03-03 grants.

Table 8. Performance results in infrared video.

Kernel	L2-regularized L2-loss (primal) C = 0.001	L2-regularized L2-loss (dual) C = 10
Mean detection time (ms)	870	924
Hits	188	186
False positives	3	6
False negatives	8	6
Accuracy (%)	94.64	94.54
Precision (%)	97.92	96.88
Recall (%)	96.91	96.88

Fig. 2. Some results in infrared video.

References

1. A.E. Delgado, M.T. López and A. Fernández-Caballero, Real-time motion detection by lateral inhibition in accumulative computation, Engineering Applications of Artificial Intelligence 23(1):129–139, 2010.
2. A. Fernández-Caballero, M.T. López, J.C. Castillo and S. Maldonado-Bascón, Real-time accumulative computation motion detectors, Sensors 9(12):10044–10065, 2009.
3. J.M. Chaquet, E.J. Carmona and A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Computer Vision and Image Understanding, `http://dx.doi.org/10.1016/j.cviu.2013.01.013`, 2013.
4. J. Moreno-Garcia, L. Rodriguez-Benitez, A. Fernández-Caballero and M.T. López, Video sequence motion tracking by fuzzification techniques, Applied Soft Computing 10(1):318–331, 2010.
5. M.T. López, A. Fernández-Caballero, M.A. Fernández, J. Mira and A.E. Delgado, Visual surveillance by dynamic visual attention method, Pattern Recognition 39(11):2194–2211, 2006.
6. A. Fernández-Caballero, J. Mira, M.A. Fernández and M.T. López, Segmentation from motion of non-rigid objects by neuronal lateral interaction, Pattern Recognition Letters 22(14):1517–1524, 2001.
7. A. Fernández-Caballero, J.C. Castillo, J. Serrano-Cuerda and S. Maldonado-Bascón, Real-time human segmentation in infrared videos, Expert Systems with Applications 38(3):2577–2584, 2011.

8. A. Fernández-Caballero, J.C. Castillo, J. Martínez-Cantos and R. Martínez-Tomás, Optical flow or image subtraction in human detection from infrared camera on mobile robot, Robotics and Autonomous Systems 58(12):1273–1281, 2010.
9. N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893, 2005.
10. S. Meysam and H. Farsi, A robust method applied to human detection, International Journal of Computer Theory and Engineering, 2(5):692–694, 2010.
11. X. Wang, T.X. Han and S. Yan, An HOG-LBP human detector with partial occlusion handling, IEEE International Conference on Computer Vision (ICCV'2009), pp. 32–39, 2009.
12. Q. Zhu, M.C. Yeh, K.T. Cheng and S. Avidan, Fast human detection using a cascade of histograms of oriented gradients, IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1491–1498, 2006.
13. J. Marin, D. Vazquez, D. Geronimo and A.M. Lopez, Learning appearance in virtual scenarios for pedestrian detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10), pp. 137–144, 2010.
14. L. Zhang, B. Wu and R. Nevatia, Pedestrian detection in infrared images based on local shape features, IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07), pp. 1–8, 2007.
15. F. Suard, A. Rakotomamonjy, A. Bensrhair and A. Broggi, Pedestrian detection using infrared images and histograms of oriented gradients, IEEE Intelligent Vehicles Symposium (IVS'06), pp. 206–212, 2006.
16. M. Bertozzi, A. Broggi, P. Grisleri, T. Graf and M. Meinecke, Pedestrian detection in infrared images, IEEE Intelligent Vehicles Symposium (IV'03), vol. 3, pp. 662–667, 2003.
17. J. Dong, J. Ge and Y. Luo, Nighttime pedestrian detection with near infrared using cascaded classifiers, IEEE International Conference on Image Processing (ICIP'07), vol. 6, pp. 185–188, 2007.
18. B.E. Boser, I.M. Guyon and V.N. Vapnik, A training algorithm for optimal margin classiers, Fifth Annual Workshop on Computational Learning Theory (COLT'92), pp. 144–152, 1992.
19. C. Cortes and V.N. Vapnik, Support-vector networks, Machine Learning, 10(3):273–297, 1995.
20. C. Papageorgiou and T. Poggio, A trainable system for object detection, International Journal of Computer Vision, 38(1):15–33, 2000.
21. R. Ronfard, C. Schmid and B. Triggs, Learning to parse pictures of people, European Conference on Computer Vision (ECCV'02), pp. 700–714, 2002.
22. D.G. Lowe, Object recognition from local scale-invariant features, IEEE International Conference on Computer Vision (ICCV'99), vol. 2, pp. 1150–1157, 1999.
23. C.C. Chang and C.J. Lin, LibSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2(3):1–27, 2011.
24. S.S. Keerthi, S. Sundararajan, K.W. Chang, C. Hsieh and C.J. Lin, A sequential dual method for large scale multi-class linear SVMs, Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 408–416, 2008.
25. R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang and C.J. Lin, LibLINEAR: a library for large linear classification, Journal of Machine Learning Research, 9:1871–1874, 2008.